ED 473 809 TM 034 791

AUTHOR	Rosenthal, Robert
TITLE	Correlations, Contrasts, and Conceptual Clarity.
PUB DATE	2002-08-00
NOTE	57p.; Paper presented at the Annual Meeting of the American Psychological Association (Chicago, IL, August 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	*Correlation; *Effect Size; Estimation (Mathematics)
IDENTIFIERS	*Pearson Product Moment Correlation

ABSTRACT

This paper discusses the Pearson product moment correlation (K. Pearson, 1986). Although this correlational metric is old, published in the 19th century, this paper suggests that it remains the most nearly universally applicable index of effect size. It is difficult to imagine a situation in which a Pearson "r" or its equivalent could not be used appropriately to index the magnitude of an event. And because all Pearson "r"s and their equivalents are based on focused comparisons, or contrasts, rather than on diffuse or omnibus comparisons, there is far greater conceptual clarity in the use of "r" than in the use of some other effect sizes measures. This paper illustrates the claims made for Pearson's "r," and discusses some applications and devices that make it more widely applicable. The paper also describes a new statistic that allows the accurate estimation of an effect size called "r" equivalent. It is also noted that there are actually four "r"s that can be usefully used as effect size estimates. Each of these is discussed. (Contains 11 tables and 52 references.) (SLD)

July 31, 2002

Correlations, Contrasts, and Conceptual Clarity

Robert Rosenthal

University of California, Riverside

Not in this century, the 21st, nor in the last, the 20th, but in the 19th century, Karl Pearson published the equation for the product moment correlation that now bears his name (Pearson, 1896; Stigler, 1986). Old as this correlational metric is, I want to suggest that it remains as perhaps the most nearly universally applicable index of effect size. It is difficult to imagine a situation in which a Pearson *r* or its equivalent, could not appropriately be used to index the magnitude of an effect. And because all Pearson *r*s, and their equivalents, are based on focused comparisons, or contrasts, rather than on diffuse or omnibus comparisons (e.g., *F* tests with more than one *df* in the numerator, or χ^2 tests with $df > 1$), there is far greater conceptual clarity in the employment of *r* than in the employment of effect size estimates based on $df > 1$.

By the use of simple displays, Pearson *r*s can be made readily understandable to policy experts who are unfamiliar with more complex statistical ideas such as standard deviation units. The interpretation of Pearson *r*s that do not differ significantly from the null value (usually zero) can be clarified by the use of simple devices like the counternull value of the obtained *r*.

In the domain of reliability of measurement, we run considerable risks when we try to get by with such non-correlational indices as percent agreement or with indices based on more than a single *df*.

In some areas of behavioral and biomedical research, effect size indices such as odds ratios and relative risks are commonly employed. It is often the case that these indices operate in ways that can be quite misleading. In such situations we can use Pearson *r*s to standardize these indices and make them more consistently useful and interpretable.

When three or more conditions are being compared in experimental or observational research, different subtypes of Pearson *r*s have been found useful. Two of these subtypes of *r*, $r_{alerting\text{-}CV}$, and $r_{contrast\text{-}CV}$, have recently been applied to the problem of construct validation permitting a useful quantification of this most complex of the types of validity of our measures.

Finally, I will describe briefly a new statistic that allows us to obtain an accurate estimate of an effect size called $r_{equivalent}$ from a knowledge only of an accurate one-tailed *p* and sample size *N*.

In the remainder of this paper I illustrate the claims I have made for Karl Pearson's 1896 invention (or discovery) and some applications and devices that make it even more widely applicable and useful. But before beginning that, I want to note that Pearson himself found a new and exciting application of his *r*. He used it in what must surely have been one of the earliest of meta-analyses.

Insert Table 1 about here

Pearson was interested in the effects of vaccination for smallpox on survival, and he collected the results of six experiments examining this relationship. Table 1 shows these six *r*s rounded to two decimal places. Pearson summarized these six correlations as an *r* of .6

Table 1 shows a few more details about his results than Pearson reported, including the mean, median, standard deviation, 95% confidence interval, one sample *t*, *p*, and $r_{contrast}$ all based on a random effects approach, i.e., treating studies (rather than patients) as sampling units (Rosenthal & DiMatteo, 2002).

Interpretive Data for Pearson's *r*

Table 2 shows four statistical procedures designed to aid in the interpretation of the correlations and their summarizers shown in Table 1. The Binomial Effect Size Display and the counternull value of an effect size can be applied to any individual correlation as well as to the mean or median *r* of a meta-analysis. The coefficient of robustness and the file drawer analysis are designed more specifically to apply to the meta-analytic context.

Insert Table 2 about here

The Binomial Effect Size Display

Table 2A shows the mean Pearson *r* of Table 1 $\left(\bar{r}=.64\right)$ as a Binomial Effect Size Display (BESD). This display is a way of showing the practical importance of any effect indexed by a correlation coefficient.The correlation is shown to be a simple difference in outcome rates between the experimental and the control groups in this standard table which always adds up to column totals of 100 and row totals of 100 (Rosenthal & Rubin, 1982). We obtain the BESD from any obtained effect size *r* by computing the treatment condition success rate as .50 plus *r*/2 and the control condition success rate as .50 minus *r*/2. Thus an *r* of .64 yields a treatment success rate of .50+.64/2 =.82 and a control success rate of .50 -.64/2 =.18. Had we been given the BESD to examine before knowing *r* we could easily have calculated it mentally for ourselves; *r* is simply the difference between the success rates of the experimental versus the control group (.82 -.18 =.64).

Pearson's mean *r* of .64 is enormous when compared to the results of most biomedical interventions. For smaller effect sizes there has been a problem in evaluating various effect size estimators from the point of view of practical usefulness (Cooper, 1981). Rosenthal and Rubin (1979; 1982) found that neither experienced behavioral researchers nor experienced statisticians had a good intuitive feel for the practical meaning of common effect size estimators and that this was particularly true for such squared indices as r^2, *omega*^{2}, epsilon^2, and similar estimates.

The Physicians' Aspirin Study. At a special meeting held on December 19,1987, it was decided to end, prematurely, a randomized double blind experiment on the effects of aspirin on reducing heart attacks (Steering Committee of the Physicians Health Study Research Group, 1988). The reason for this unusual termination of such an experiment was that it had become so clear that aspirin prevented heart attacks (and deaths from heart attacks) that it would be unethical to continue to give half the physician research subjects a placebo. And what was the magnitude of the experimental effect that was so dramatic as to call for the termination of this research? Was r^2 .80 or .60, so that the corresponding *r*s would have been .89 or .77? Was r^2 .40 or .20, so that the corresponding *r*s would have been .63 or .45? No, none of these. Actually r^2 was .00 or, to four decimal places, .0011, with a corresponding *r* of .034. The decision to end the aspirin experiment was an ethical necessity -- it saved lives. Most social and behavioral scientists are surprised that life-saving interventions can be associated with effect sizes as small as *r*s of .034 and r^2s of .0011.

This type of result seen in the Physicians' Aspirin Study is not at all unusual in biomedical research. Some years earlier, on October 29, 1981, the National Heart, Lung, and Blood Institute discontinued its placebo-controlled study of propranolol because results were so favorable to the treatment that it would be unethical to continue withholding the life-saving drug from the control patients (Kolata, 1981). Once again the effect size r was .04, and the leading digits of the r^2 were .00! As behavioral researchers we are not used to thinking of *r*s of .04 as reflecting effect sizes of practical importance. But when we think of an r of .04 as reflecting a 4% decrease in heart attacks, the interpretation given r in a Binomial Effect Size Display, the r does not appear to be quite so small.

The Counternull Value of an Effect Size

Table 2B shows the counternull value of Pearson's mean r. The counternull was recently introduced as a new statistic (Rosenthal & Rubin, 1994). It is useful in virtually eliminating two common errors: (a) equating failure to reject the null with the estimation of the effect size as equal to zero and (b) equating rejection of a null hypothesis on the basis of a significance test with having demonstrated a scientifically important effect. In most applications, the value of the counternull is simply twice the magnitude of the obtained effect size (e.g., Cohen's d, Hedges's g, Glass's Δ, Z_r). Thus with $r = .10$ found to be nonsignificant, the counternull value of $r = .20$ is exactly as likely as the null value of $r = .00$. For any effect size with a symmetric reference distribution such as the normal or any t distribution, the counternull value of an effect size can always be found by doubling the obtained effect size and subtracting the effect size expected under the null hypothesis (usually zero). Thus, if we found that a test of significance did not reach the chosen level (e.g., .05), the use of the counternull would keep us from concluding that the mean effect size was, therefore, probably zero. The counternull value of $2d$ or $2Z_r$

would be just as tenable a conclusion as concluding $d=0$ or $Z_r=0$. In our example of Pearson's meta-analysis, the counternull value of Z_r was 1.52 and, therefore, in units of r the counternull value was .91, an extremely large value.

The counternull is a kind of confidence interval conceptually related to the more traditional (e.g., 95%) confidence interval. As Cohen, with his customary wisdom, pointed out, the behavioral and medical sciences would be more advanced had we always routinely reported not just p values but effect size estimates with confidence intervals as well (Cohen, 1990; 1994).

The Coefficient of Robustness

The standard error of the mean effect size in a meta-analysis, along with confidence intervals placed around the mean effect size are of great value (Rosenthal & Rubin, 1978). It will often be useful also to employ a statistic that does not change simply as a function of the increasing number of replications. Thus, if we want to compare two research areas for their robustness, adjusting for the difference in number of replications in each research area, we may prefer the "robustness coefficient". For example, Table 2C shows the coefficient of robustness for the smallpox study to be 8.89, a value we can compare to another of Pearson's meta-analyses, that of the survival value of inoculation against typhoid. That coefficient of robustness was a more modest, but still impressive, 1.58.

The coefficient of robustness is simply the mean effect size divided by the S of the effect sizes. This metric is the reciprocal of the coefficient of variation (Rosenthal, 1990; 1993). The coefficient of robustness (*CR*) can also be viewed in terms of the one-sample *t* test on the mean of the set of *k* effect sizes, when each is given equal weight. Thus, *CR* is given by $t/\sqrt{k}$, or *t* adjusted for number of studies.

The utility of this coefficient is based on two ideas. First, robustness (or replication success, or clarity) depends on the homogeneity of the obtained effect sizes. Second, robustness depends also on the unambiguousness or clarity of the directionality of the result. Thus, a set of replications grows in robustness when the variability (*S*) of the effect sizes (the denominator of the coefficient) decreases and also when the mean effect size (the numerator of the coefficient) increases. Incidentally, the mean may be weighted, unweighted, or "trimmed" (Tukey, 1977). Indeed, it need not be the mean at all, but any measure of location or central tendency (e.g., the unweighted or weighted median).

The coefficient of robustness can be seen as a kind of second order effect size. An illustration will be helpful. Imagine that three meta-analyses of three treatments have been conducted with mean effect size *d*s of .8, .6, and .4, respectively. If the variability (*S*) of the three meta-analyses were quite similar to one another, the analysis showing the .8 mean *d* would, of course, be declared the most robust. However, suppose the *S*s for the three analyses were 1.00, 0.60, and 0.20, respectively. Then the three coefficients of robustness would be .8/1.00 = .8, .6/.60 = 1.0, and .4/.20 = 2.0. Assuming reasonable and comparable sample sizes and numbers of studies collected for the three analyses, the treatment with the *smallest* effect size (i.e., .4) would be declared most robust with the implication that its effect is the most consistently positive.

The File Drawer Analysis

The *file drawer problem* refers to the well-supported suspicion that the studies retrievable in a meta-analysis are not likely to be a random sample of all studies actually conducted (Rosenthal, 1979; 1991a). The suspicion has been that studies actually published are more likely to have achieved statistical significance than the studies remaining squirreled away in the file drawers (Sterling, 1959). No definitive solution to this problem is available, but we can establish reasonable boundaries on the problem and we can estimate the degree of damage to any research conclusion that could be done by the file drawer problem. The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error can be brought to any precisely specified level of significance, say $p = .05$. This number of filed studies, or the *tolerance for future null results*, is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the reviewer. If the overall level of significance of the research review will be brought down to the level of *barely significant* by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

Details of the (fixed effect) calculations and rationale are given elsewhere (Rosenthal, 1991a), but briefly, for a random effects analysis based on *k* studies, we can find the number (X) of new, filed, or unretrieved studies averaging null results (i.e., r =.00) required to bring the new overall p to .05 from the following:

$$X = \left[\frac{(\Sigma r)^2}{(2.706)S_r^2} \right] - k \qquad (1)$$

where Σr is the sum of the rs of all the k studies we have retrieved, and S_r^2 is the variance of the rs of the k retrieved studies. Table 2D shows that over 1,000 unretrieved studies averaging null results would be required to bring the overall level of significance to .05 or greater. By way of comparison, in Pearson's typhoid meta-analysis, only 27 such studies would be required.

It should be noted that the file drawer analysis addresses only the effects on publication bias of the results of significance testing. Very sophisticated graphic (Light & Pillemer, 1984), and other valuable procedures are available for the estimation and correction of publication bias (e.g., Begg, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 1990).

Risks in Not Using Pearson's *r*-Based Indices of Reliability

Percent Agreement

It has long been common practice for some researchers to index the reliability of judges' categorizations using percent agreement defined as

$$\left(\frac{A}{A + D} \right) 100 \qquad (2)$$

where A represents the number of agreements and D represents the number of

----

Insert Table 3 about here

----

disagreements (Rosenthal & Rosnow, 1991).

Table 3 shows how percent agreement can be a very misleading indicator of interjudge reliability. In Part A of Table 3 we find that two researchers, Smith and Jones, each had two judges evaluate a series of 100 film clips of children for the presence or absence of frowning behavior. Both Smith and Jones found their judges to show 98% agreement , but Smith's 98% agreement was a hollow victory indeed. The correlation between Judges A and B was actually slightly negative, $r = -.01$, $\left(\chi^2_{(1)} = 0.01\right)$. Jones's 98% agreement, on the other hand, was associated with an r of $+.96$, $\left(\chi^2_{(1)} = 92.16\right)$.

Part B of Table 3 shows two additional cases of percent agreement obtained by researchers North and West. This time, the two investigators have both obtained an apparently chance level of agreement, i.e., 50%. Both results, however, are very far from reflecting chance agreement, both with $p = .0009$. Most surprising, perhaps, is that North obtained a substantial negative reliability (r= -.33) while West obtained a substantial positive reliability (r = + .33); another illustration that percent agreement is not a very informative index of reliability.

Multi-df Interjudge Reliability

Among the first psychologists to appreciate the problems of percent agreement as an index of reliability was Jacob Cohen (1960). He developed an index, *kappa*, that solved the problem of the percent agreement index by adjusting for any agreement based simply on lack of variability, e.g., the lack of variability found in Table 3A where both of Smith's judges found 99% of the film clips to show frowning behavior.

Insert Table 4 about here

Table 4 gives an example of the type of situation in which *kappa* is often employed. Two clinical diagnosticians have examined 100 people and assigned them to one of four classifications, e.g., schizophrenic, neurotic, normal, and brain damaged. Only three quantities are required to compute *kappa*:

O = observed number on which the two judges have agreed, i.e., the number on the diagonal of agreement; in this example: $13+12+12+13 = 50$.

E = expected number under the hypothesis of only chance agreement for the cells on the diagonal of agreement. For each cell, the expected number is the product of the row total and the column total divided by the total number of cases. In this example the expected number is:
$(25 \times 25)/100 + (25 \times 25)/100 + (25 \times 25)/100 + (25 \times 25)/100 =$
$6.25 + 6.25 + 6.25 + 6.25 = 25.$

N = total number of cases classified; in this example, $N=100$.

kappa is computed from

$$kappa = \frac{O-E}{N-E} = \frac{50-25}{100-25} = .333, \quad (3)$$

in the present example.

Although *kappa* is clearly an improvement over percent agreement as an index of reliability, it does raise some serious questions. When *kappa* is based on tables larger than a 2 x 2, e.g., a 3 x 3, a 4 x 4 (as in Table 4), or larger, as it often is, *kappa* suffers from the same problem as does any statistic on *df*>1. That problem, the problem of diffuse or omnibus procedures, is that for most values of *kappa* we cannot tell which focused or specific judgments are made reliably and which are made unreliably. Only when *kappa* approaches unity is the actual interpretation of a value of *kappa* straightforward, i.e., essentially all judgments are made reliably (Rosenthal, 1991b). We illustrate the difficulty in interpreting *kappa* by returning to Table 4.

The 4 x 4 table we see, based on 9 *df,* can be decomposed into a series of six pairwise 2 x 2 tables each based on a single *df*, and addressing a very specific, conceptually clear question of the reliability of dichotomous judgments; A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D. Table 5 shows the results of computing *kappa* separately for each of these six 2 x 2 tables.

Insert Table 5 about here

Of the six focused or specific reliabilities computed, four are *kappas* of 1.00, and two are *kappas* near zero (.04 and -.04). The mean of the six 1 *df kappas* is

.667, and the median is 1.00; neither value being predictable from the omnibus 9 *df kappa* value of .33. To show even more clearly how little relation there is between the omnibus values of *kappa* and the associated 1 *df kappas*, i.e., the focused reliability *kappas*, Tables 6 and 7 have been prepared. Table 6 shows an omnibus 9 *df kappa* value of .33, exactly the same value as that shown in Table 4.

Insert Tables 6 and 7 about here

Table 7 shows the six focused reliabilities of *df*=1 associated with the omnibus value of *kappa* (.33) of Table 6. We see that of these six focused *kappas*, four are *kappas* of .00, one is a *kappa* of +1.00, and one is a *kappa* of -1.00. The mean and median focused *kappa* both show a value of .00. We can summarize the two omnibus *kappas* of Tables 4 and 6 and their associated focused *kappas* as follows:

	Example 1	Example 2
Omnibus *kappa*	.33	.33
Mean focused *kappa*	.67	.00
Median focused *kappa*	1.00	.00

Thus we have two identical *kappas*; one made up primarily of perfect reliabilities, the other made up primarily of zero reliabilities.

Although the greatest limitations on *kappa* occur when *kappa* is based on *df*>1, there are some problems with *kappa* even when it is based on a 2 x 2 table of counts where *df*=1. The basic problem under these conditions is that very often *kappa* is not equivalent to the product moment correlation computed from exactly the same 2 x 2 table of counts. This is certainly not a criticism of *kappa* since it never pretended to be a product moment correlation. The limitation, however, is that we cannot apply various interpretive procedures or displays to *kappa* that we can apply to product moment correlations. Examples include the use of the *coefficient of determination* (i.e., r^2) and the *Binomial Effect Size Display*.

Here we need only indicate the conditions under which a 1 *df kappa* is or is not equivalent to a product moment correlation (referred to as a Pearson *r* in the general case and sometimes referred to as *phi* - or ϕ - in the case of a 2 x 2 table of counts). *Kappa* and *r* are equivalent when the row totals for levels A and B are identical to the column totals for levels A and B, respectively. Consider the following example:

		Judge 1		
		A	B	Σ
	A	70	10	80
Judge 2	B	10	10	20
	Σ	80	20	100

For these data, where the marginal totals for level A are identical for Judges 1 and 2 (i.e., 80),

$$kappa(df = 1) = \frac{O-E}{N-E} = \frac{80-68}{100-68} = .375,$$

and *r* (or equivalently, *phi*) yields the identical value of .375. Therefore, we could meaningfully compute a coefficient of determination or a *Binomial Effect Size Display* for this particular *kappa* because it is equivalent to a Pearson *r* or *phi* (ϕ).

Now consider the following example in which we have the same four cell entries and the same marginal totals as in the preceding example. The only thing that has changed is the location of the cell with the largest count (70) so that the marginal totals for level A differ for Judges 1 and 2 (20 versus 80).

		Judge 1		
		A	B	Σ
	A	10	70	80
Judge 2	B	10	10	20
	Σ	20	80	100

In this example,

$$kappa(df=1) \;=\; \frac{O-E}{N-E} \;=\; \frac{20-32}{100-32} \;=\; -.176,$$

but *r* (or ϕ) yields a markedly different value of -.375. We can, therefore, compute a meaningful coefficient of determination and *Binomial Effect Size Display* for *r*, but we cannot do so for *kappa*.

Using Pearson's *r* to Improve Other Effect Size Estimates:
2 x 2 Tables of Counts in the Biomedical Context

The effect size index, *r*, can, of course be readily applied to any 2 x 2 table of counts. Three other indices of effect size have also been frequently employed, especially in biomedical contexts. These are: (a) relative risk, (b) odds ratio, and (c) risk difference. All three are illustrated for several hypothetical outcomes in Table 8. Each study compared a control condition to a treatment condition with two possible outcomes: not surviving or surviving.

Insert Table 8 about here

Relative Risk

Relative risk is defined as the ratio of the proportion of the control patients at risk (not surviving) divided by the proportion of the treated patients at risk. With the cells of the 2 x 2 table of counts labeled A, B, C, and D from upper left to lower right (as shown in Table 8) relative risk (RR) is defined as:

$$RR = \left(\frac{A}{A+B} \Big/ \frac{C}{C+D} \right)$$

A limitation of this effect size estimate can be seen in Table 8. We examine the three study outcomes closely and ask ourselves the following: If we had to be in the control condition would it matter to us whether we were in Study I, Study II, or Study III? We think most people would rather have been in Study I than II and we think that virtually no one would have preferred to be a member of the control group in Study III. Yet, despite the very important phenomenological differences among these three studies, Table 8 shows that all three relative risks are identical: 10.00. That feature may be a serious limitation to the value and informativeness of the relative risk index.

Odds Ratio

The odds ratio is defined as the ratio of the not surviving control patients to the surviving control patients divided by the ratio of the not surviving treated patients to the surviving treated patients. For cells as labeled in Table 8, the odds ratio (OR) is defined as:

$$OR = \left(\frac{A}{B}\Big/\frac{C}{D}\right)$$

The odds ratio behaves more as expected in Table 8 than does the relative risk in that the odds ratio increases with our phenomenological discomfort as we go from the results of Study I to Study II to Study III. But the high odds ratio for Study I seems alarmist. Indeed, if the data showed:

	Die	Live	
Control	10	999,990	10^6
Treated	1	999,999	10^6
	11	1,999,989	$2(10^6)$

so that an even smaller proportion of patients were at risk, the odds ratio would remain at 10.00, an even more alarmist result.

The odds ratio for Study III is also unattractive. Since all the controls die, perhaps we could forgive the infinite odds ratio. However, very different phenomenological results yield an identical odds ratio. If the data showed:

	Die	Live	
Control	1,000,000	0	10^6
Treated	999,999	1	10^6
	1,999,999	1	$2(10^6)$

we would again have an infinite odds ratio, definitely an alarmist result. In this case even the problematic relative risk index would yield a phenomenologically more realistic result of 1.00.

Risk Difference

The risk difference is defined as the difference between the proportion of the control patients at risk and the proportion of the treated patients at risk. For cells as labeled in Table 8, the risk difference (RD) is defined as:

$$RD = \left(\frac{A}{A+B} - \frac{C}{C+D}\right)$$

The last column of Table 8 shows the Pearson product moment correlation (*r*) between the independent variable of treatment (scored 0,1) and the dependent variable of outcome (scored 0, 1). Comparison of the risk differences with *r* in Table 8 (and elsewhere) shows that the risk difference index is never unreasonably far from the value of *r*. For that reason the risk difference index may be the one least likely to be quite misleading under special circumstances and so we prefer it as our all-purpose index if we had to use one of the three indices under discussion. But even here we feel we can do better.

Standardizing the Three Risk Measures

We propose a simple adjustment that standardizes our measures of relative risk, odds ratio, and risk difference (Rosenthal, 2000; Rosenthal, Rosnow, & Rubin, 2000; Rosenthal & Rubin, 1998). We simply compute the correlation *r* between the treatment and outcome and display *r* in a Binomial Effect Size Display (BESD) as described earlier.

Insert Table 9 about here

Table 9 shows the BESD for the three studies of Table 8. Although the tables of counts of Table 8 varied from *N*s of 2,000, to 40, to 20, the corresponding BESDs of Table 9 all show the standard margins of 100 which is a design feature of the BESD. The computation of our new effect size indices is straightforward. We simply compute relative risks, odds ratios, and risk differences on our standardized tables (BESDs) to obtain standardized relative risks, standardized odds ratios, and standardized risk differences. The computation of these three indices is simplified because the A and D cells of a BESD always have the same value (as do the B and C cells). Thus the computational equations simplify to A/C for standardized relative risk (SRR), to $(A/C)^2$ for standardized odds ratio (SOR), and to (A-C)/100 for standardized risk difference (SRD).

Table 9 shows the standardized relative risks increasing as they should in going from Study I to Study III. The standardized odds ratios also increase as they go from Study I to Study III but without the alarmist value for Study I and the infinite value for Study III. (A standardized odds ratio could go to infinity only if *r* were exactly 1.00, an unlikely event in behavioral or biomedical research.) The standardized risk difference is shown in Table 9 to be identical to *r* which is an attractive feature emphasizing the interpretability of *r* as displayed in a BESD.

One Pearson's *r*; Four Useful Subtypes

So far in our joyful praise of *r* as a highly valued effect size estimate, we have not mentioned the fact that there are actually *four* *r*s that can be usefully employed as effect size estimates. That is the case both in meta-analytic work and in the analysis of the data of a single study. The effect size *r* most often employed is only one of those *r*s, specifically, $r_{contrast}$. Ideally, both in meta-analytic work and in the analysis of the data of individual studies, we would report all four correlations, because each of them addresses a different question (Rosenthal, Rosnow & Rubin, 2000).

$r_{contrast}$

This *r* is a partial correlation between individual sampling units' scores on the dependent variable and the predicted mean score (contrast weight) of the group to which they belong -- with other between group variation partialed out. This is the most frequently used correlation in meta-analytic work because it is often the only correlation we can calculate from other people's data. We can find $r_{contrast}$ from tests of significance by any of the following equations:

$$r_{contrast} = \sqrt{\frac{F_{contrast}}{F_{contrast} + df_{within}}}\ , \tag{4}$$

$$r_{contrast} = \sqrt{\frac{t^2_{contrast}}{t^2_{contrast} + df_{within}}}\ , \tag{5}$$

$$r_{contrast} = \sqrt{\frac{\chi^2_{(1)}}{N}}\ , \tag{6}$$

$$r_{contrast} = \frac{Z}{\sqrt{N}}\ . \tag{7}$$

and we can compute $r_{contrast}$ from the effect size estimate d from the following:

$$r_{contrast} = \sqrt{\frac{d^2}{d^2 + 4}}\ , \tag{8}$$

For further details on other equivalences among effect size estimates see Rosenthal, 1991a, 1994; and Rosenthal and Rosnow, 1991.

In the simplest case, where two groups are being compared, $r_{contrast}$ is the point biserial correlation between membership in one of the two groups (coded, e.g., 0 and 1) and the score on the dependent variable. In this simple two-group case we report only the value of $r_{contrast}$ and not the values of the other three correlations.

When there are three or more groups being studied, however, each of the four correlations tells us something different about the relationship between the independent and dependent variable. For example, $r_{alerting}$, the correlation between the predicted and obtained mean scores per condition, often alerts us to an otherwise overlooked relationship. For example, we may read a report saying there is "no relationship" between age level (e.g., ages 8, 9, 10, 11, 12) and cognitive performance with $F_{(4,\ 20)} = 1.50$, $p = .24$. However, looking at the five means of this report may show a perfect correlation ($r_{alerting}$) between age level and mean performance, clearly contradicting the conclusion of the report that there was no relationship between age and performance. That claim had been based on an inappropriate omnibus *F* test with 4 *df* in the numerator. A properly computed $F_{contrast}$ would have yielded $F_{(1,\ 20)} = 6.00$, $p = .024$, $r_{contrast} = .48$, ($r_{alerting}$ $= 1.00$, *t* very large, *p* very small). Other uses of $r_{alerting}$ include its role in the computation of contrasts in other people's data (Rosenthal & Rosnow, 1985; Rosnow & Rosenthal, 1996; Rosnow & Rosenthal, in press; Rosnow, Rosenthal, & Rubin, 2000).

$r_{effect\ size}$

This is the correlation between individual sampling units' scores on the dependent variable and the predicted mean score (contrast weight) of the group to which they belong without any partialing. $r_{effect\ size}$, because it involves no partialing of other between group effects out of the error term, is never larger than $r_{contrast}$ and is usually smaller than $r_{contrast}$; sometimes dramatically so. $r_{effect\ size}$ can be computed from

$$r_{effect\ size} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}\left(df_{noncontrost}\right) + df_{within}}} \quad . \qquad (9)$$

$r_{alerting}$

This is the correlation between the condition means and the predicted mean scores (contrast weights). $r_{alerting}$ can be computed from

$$r_{alerting} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}\left(df_{noncontrast}\right)}} \quad . \tag{10}$$

r_{BESD}

This is a usually more conservative effect size correlation that permits generalization not only to other sampling units in the same conditions but also to other levels of the same independent variable. r_{BESD} can be computed from

$$r_{BESD} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}\left(df_{noncontrast} + df_{within}\right)}} \quad . \tag{11}$$

In Equation 11, just above, when $F_{noncontrast}$ is less than 1.00 it is entered in Equation 11 as equal to 1.00. $F_{noncontrast}$ is computed as

$$\frac{F_{between}\left(df_{between}\right) - F_{contrast}}{df_{between} - 1} \quad . \tag{12}$$

The restriction that $F_{noncontrast}$ in Equation 11 cannot drop below 1.00 formalizes the assumption that the noncontrast variation is noise and forces r_{BESD} to be less than, or at most equal to, $r_{effect\ size}$. Detailed discussions of these four correlations are provided in Rosenthal, Rosnow, and Rubin, (2000).

Using Pearson's *r* to Quantify Construct Validity

Construct validity is one of the most important concepts in all of psychology. Yet despite the importance of this concept, no simple metric can be employed to quantify the extent to which a measure can be described as construct valid. Researchers typically establish construct validity by presenting correlations between a measure of a construct and a number of other measures that should, theoretically, be associated with it (convergent validity) or vary independently of it (discriminant validity).

The aim of construct validation is to embed a purported measure of a construct in a nomological network, that is, to establish its relation to other variables with which it should, theoretically, be associated positively, negatively, or practically not at all (Cronbach and Meehl, 1955). A procedure designed to help quantify construct validity should provide a summary index not only of *whether* the measure correlates positively, negatively, or not at all with a series of other measures, but the relative magnitude of those correlations. Or put another way, it should be an index of the extent to which the researcher has accurately predicted the pattern of findings in the convergent-discriminant validity array. Such a metric should also provide a test of the statistical significance of the match between observed and expected correlations, and provide confidence intervals for that match, taking into account the likelihood that some of the validating variables may not be independent of one another.

In a recent paper, Drew Westen and I present two effect size estimates (both Pearson *r*s) for quantifying construct validity (Westen & Rosenthal, 2002). These two *r*s, variants on two of the four *r*s described in the previous section, were designed to summarize the pattern of findings represented in a convergent-discriminant validity matrix for a given measure. These metrics provide simple estimates of validity that can be compared across studies, constructs, and measures. Both metrics provide a quantified index of the degree of convergence between the observed pattern of correlations and the theoretically predicted pattern of correlations -- that is, of the degree of agreement of the data with the theory underlying the construct and the measure.

Contrasts and Construct Validity

In their classic paper on construct validation, Cronbach and Meehl (1955) considered the possibility of developing an overall coefficient for indexing construct validity but noted the difficulty of providing anything more than a broad indication of the upper and lower bounds of validity. However, developments since that time, particularly in the concept of the multi-trait multi-method matrix (MTMM) (Campbell and Fiske, 1959; Shrout and Fiske, 1995), have led to continued efforts to derive more quantitative, less impressionistic ways to index the extent to which a measure is doing its job. Thus, a number of researchers have developed techniques to try to separate out true variance on a measure of a trait from method variance, often based on the principle that method effects and trait effects (and their interactions) should be distinguishable using analysis of variance, confirmatory factor analysis (because trait and method variance should load on different factors), structural equation modeling, and related statistical procedures (Cudeck, 1988;

Hammond, Hamm, & Grassia, 1986; Kenny, 1995; Reichardt and Coleman, 1995; Wothke, 1995).

Our procedures are in many respects related, but are simple, readily applied, and designed to address the most common case in which a researcher wants to validate a single measure by correlating it with multiple other measures.

The approach we proposed, based on contrast analysis, asks a highly specific, focused question with one degree of freedom. The question it addresses is whether the researcher has accurately predicted the magnitude of correlations between a single predictor variable and multiple criterion variables. Rosenthal, Rosnow, & Rubin (2000) have outlined the advantages of focused questions of this sort, but the major advantage is that these procedures, based on one degree of freedom, provide a single answer to a single question; in this case, does this measure predict an array of correlations with other measures in a way predicted by theory?

The procedures Drew Westen and I proposed derive primarily from recent developments in contrast analysis (Meng, Rosenthal, & Rubin, 1992; Rosenthal, Rosnow, & Rubin, 2000), a set of techniques usually employed in the analysis of variance to test specific hypotheses about the relative magnitude of a series of means. Although researchers have most commonly applied this method to analysis of variance in experimental designs, contrast analysis is equally applicable to correlational data. Just as researchers can construct contrasts to test the relative ordering of means, they can equally construct contrasts to assess the relative ordering of correlation coefficients, even when those correlation coefficients are correlated with one another (Meng, Rosenthal, & Rubin, 1992; Rosenthal, Rosnow, & Rubin, 2000).

Two Pearson rs for Construct Validity: $r_{alerting\text{-}CV}$ and $r_{contrast\text{-}CV}$

Two Pearson *r*s provide convenient and informative indices of construct validity, each in its own way. The first of these correlations, $r_{alerting\text{-}CV}$, is the simple correlation between (a) the pattern of correlations *predicted* between the measure being validated and the *k* variables correlated with that measure, and (b) the pattern of correlations actually *obtained*. It is called an "alerting" correlation because it is a rough, readily interpretable index that can alert the researcher to possible trends of interest (Rosenthal et al., 2000).

For example, suppose we were developing a new measure of interpersonal skill. We have administered our new measure to a sample of participants to whom we have also administered four other measures. Our construct of interpersonal skill is such that we predict it will correlate with the four other measures as follows: (1) Verbal IQ, *r* predicted roughly as .5, (2) Nonverbal decoding skill, *r* predicted roughly as .5, (3) Agreeableness, *r* predicted roughly also as .5, and (4) Conscientiousness, *r* predicted as .1. To compute $r_{alerting\text{-}CV}$ we simply correlate these predicted values (arranged as a column of data) with the obtained values (arranged as a second column of data). More accurate results are obtained when the correlations (*r*s) are first transformed into their Fisher Z_r equivalents in order to improve normality (Meng, et al., 1992; Steiger, 1980).

Thus, suppose the obtained values, Z_r transformed, were .74, .59, .60, and -.03. The correlation between this column of data and our predicted values

(.5, .5, .5, .1) yields an $r_{alerting\text{-}CV}$ of .98. The magnitude of this correlation suggests that our predicted pattern of values provided a very accurate portrayal of the pattern or profile of correlations actually obtained.

The effect size correlation $r_{alerting\text{-}CV}$ becomes increasingly useful as we include more and more variables in our convergent-discriminant validity matrix. If only two variables are to be correlated with our new measure, $r_{alerting\text{-}CV}$ can take on values of only +1.00 or -1.00. As more variables are added, $r_{alerting\text{-}CV}$ becomes more informative. To put it another way, $r_{alerting\text{-}CV}$ provides an unstable index when the number of criterion variables is small but becomes progressively more useful as the researcher makes bolder hypotheses about the relation between the target measure and a range of criterion variables—that is, as the nomological net gets wider. We typically do not compute *p* levels for $r_{alerting\text{-}CV}$, but it can be used to help in the computation of significance levels for our other effect size correlation $r_{contrast\text{-}CV}$.

Our second correlation, $r_{contrast\text{-}CV}$, shares with $r_{alerting\text{-}CV}$ the characteristic that it will be larger as the match between expected and obtained correlations is higher. In addition, however, $r_{contrast\text{-}CV}$ uses information about, (a) the median intercorrelation among the variables to be correlated with the measure being validated, and (b) the absolute values of the correlations between the measure being validated and the variables with which it is being correlated. A desirable feature of $r_{contrast\text{-}CV}$ is that its interpretation is not limited in the same way as is $r_{alerting\text{-}CV}$ when there are only a few variables in the convergent-discriminant validity matrix. Computational details for $r_{contrast\text{-}CV}$ are provided in Appendix A of Westen and Rosenthal, (2002), and in a less directly applicable form, in Meng, et al, (1992).

Insert Table 10 about here

Table 10 shows the intercorrelations among our five variables, including the new measure we are in the process of validating, and the four variables for which we have predicted the correlations with the new measure that would contribute to its construct validation. We have already reported $r_{alerting\text{-}CV}$ as .98; we now report $r_{contrast\text{-}CV}$ to be .60. The equations given in Westen and Rosenthal (2002) and in Meng, et al. (1992) also yield a χ^2 (on *k*-1 *df*) testing the heterogeneity of the set of correlations of the validating variables with the common dependent variable (i.e., the new measure). For the data of Table 10, this $\chi^2(3) = 5.71$. Interestingly, the *Z* test of significance of $r_{contrast\text{-}CV}$ can be obtained by multiplying $r_{alerting\text{-}CV}$ by the square root of the χ^2 test for heterogeneity; in this example,

$$Z = r_{alerting-CV}\sqrt{\chi^2(k-1)} = (.98)\ \sqrt{5.71} = 2.34,\ p = \ .0096. \qquad (13)$$

We can get $r_{contrast\text{-}CV}$ from Equation (7) yielding

$$r_{contrast-CV} = \frac{2.34}{\sqrt{15}} = \ .604$$

or from Equation (5) employing *t* instead of *Z*. We get *t* from the *p* associated with *Z* (.0096 in this case) and therefore find $t_{(13)}$ to be 2.67. Then from Equation (5) we find

$$r_{contrast-CV} = \sqrt{\frac{(2.67)^2}{(2.67)^2 + 13}} = .595,$$

a value slightly lower than the .604 obtained from Equation (7). With large samples Equations (5) and (7) tend to give the same values; with smaller sample sizes Equation (5) employing *t*, tends to more accurate.

Getting Pearson's *r* from *p*: $r_{equivalent}$

Recent years have shown increasing dissatisfaction with the use of dichotomous decision-making based on significance tests and an increased recognition of the value of reporting effect sizes. Indeed, the report of the Task Force on Statistical Inference of the Board of Scientific Affairs of the American Psychological Association has explicitly recommended that the primary results of any research should be presented as effect sizes, preferably with an accompanying confidence interval (Wilkinson & the Task Force on Statistical Inference, 1999).

The purpose of a recent paper by Don Rubin and myself is to describe a simple procedure for obtaining an accurate estimate of an effect size from a *p*-value and the sample size (Rosenthal & Rubin, 2002). This procedure is especially appropriate when:

1. In meta-analytic work, or in other re-analyses of others' studies, neither effect sizes nor significance tests are provided, but only *p*-values and sample sizes are reported,

2. No effect size estimate has been generally accepted for the data analytic procedures employed, or

3. An effect size estimate can be computed directly from the data but, because of small sample sizes or severe nonnormality, the estimates may be seriously misleading.

Meta-Analytic Research in Which Only p-Values Have Been Reported

In conducting meta-analyses we often find that only *p*-values have been provided rather than effect size estimates or significance test statistics such as *t* or *Z*, or one *df F* or χ^2. When those *p*-values are reported accurately, e.g., $p = .11$, $p = .02$, $p = .003$, we can get accurate effect size estimates from them and the sample size. When *p*s are reported only as $<.05$, $<.01$, etc., we cannot get accurate effect size estimates but we can set lower bounds, i.e., the lowest possible value of the effect size, but not upper bounds, the highest possible value of the effect size. The fact that in meta-analytic applications we can sometimes obtain only lower bound values must be kept in mind, but such lower bound, conservative estimates of effect size are better than having no estimate at all.

No Generally Accepted Effect Size Estimate Exists

Many effect size estimates have been described and have been widely used (e.g., Cohen, 1988, Fleiss, 1994; Rosenthal, 1991; 1994). However, there remain numerous statistical procedures for which no standard effect size estimate is recognized, for example, for many distribution-free or nonparametric procedures. What effect size estimate should we use, for example, when we have computed *p*-values from Fisher's exact test, or from a sign test, a one sample runs test, a Wilcoxon signed ranks test, a Mann-Whitney U test, or other permutation tests (Siegel & Castellan, 1988)?

Directly Computed Effect Size Estimates are Likely to be Seriously Misleading

Consider a very small randomized experiment in which three animals are vaccinated and all survive, and three animals are not vaccinated and do not survive. The sample correlation between vaccination and survival for these six animals is +1.00. Because of the small sample size and the nonnormality of survival, the obtained sample correlation is probably a very misleading estimate of the population correlation. We can do better by computing an accurate *p* for these six animals and then using *p* to compute a more appropriate effect size estimate $r_{equivalent}$.

Computing $r_{equivalent}$

Our procedure yields $r_{equivalent}$ from an accurate one-tailed *p* and sample size *N* by obtaining the value of *t* (with $df = N\text{-}2$) associated with the one-tailed *p*-value. One-tailed *p*s in the "wrong" or unpredicted direction are recorded as $r_{equivalent}$ with a negative sign. We find these values of *t* quite readily from extended tables of *t*, from hand held calculators, or from computers. Once we have the *t* associated with the one-tailed *p* and *N*, we compute $r_{equivalent}$ from:

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2 + (N-2)}} \quad , \qquad (14)$$

a well-known general relationship (shown earlier as Equation 5; Cohen, 1965; Rosenthal & Rosnow, 1991). When the *p*-value we used to obtain the value of *t* was based on a contrast employing more than two conditions, we replace the expression (*N* -2) in Equation (14) by the expression (*N* - *k*) where *k* is the number of conditions. Even more generally, *N*-2 is replaced by the degrees of freedom on which the *p*-value is based.

The interpretation of $r_{equivalent}$ is that it is the sample point-biserial correlation we would have found in data yielding our obtained *p*-value in a two group, equal *n* study with *N*/2 in each group. Although technically we assume that the data exactly met the usual assumptions required for the *t* test (iid normal, with the same variance in each group), $r_{equivalent}$ can be a very useful approximate effect size estimator even when these assumptions are not met precisely.

That is, suppose we conducted a randomized experiment with *N*/2 assigned to the treatment condition and N/2 assigned to the control condition. Also suppose that the data are independently normally distributed in each condition with the same variance. Then, when the value of the *t*-test statistic is *t* with the obtained *p*-value, the value of the point biserial correlation between treatment condition and outcome is $r_{equivalent}$.

Confidence Intervals for $r_{equivalent}$

More research is needed to set appropriate confidence intervals for $r_{equivalent}$. Until that research becomes available, however, we believe the usual procedure for forming confidence intervals will work well for $r_{equivalent}$. Thus, a 95% confidence interval around the Fisher *Z* transformed $r_{equivalent}$ can be found from Equation (15).

$$95\%\ CI = Z_r \pm 1.96/\sqrt{N-3}\ . \qquad (15)$$

A Simple Example

Earlier we described a randomized experiment in which three vaccinated animals survived and three unvaccinated animals did not survive, yielding a sample correlation of 1.00 between being vaccinated and survival. We can obtain an accurate *p*-value for these data from Fisher's exact test:

$$p = \frac{3!3!3!3!}{6!3!0!0!3!} = .05, \text{ one-tailed.}$$

Hence $p = .05$ and $N = 6$, so $t_{(4)} = 2.13$, and from Equation (14) we find:

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2+(N-2)}} = \sqrt{\frac{(2.13)^2}{(2.13)^2+(6-2)}} = .73\ ,$$

a more realistic estimate of the population value of the correlation between vaccination and survival than the estimate of 1.00 based on the correlation in the sample.

We now use Equation (15) to compute a 95% confidence interval around the obtained $r_{equivalent}$. For $r_{equivalent} = .73$, we find $Z_r = .93$ so, with $N = 6$, the 95% CI around Z_r runs from $.93 - 1.96/\sqrt{3}\ to\ .93 + 1.96/\sqrt{3}$ or from -.20 to +2.06. Transforming our 95% CI for Z_r back to a 95% CI for $r_{equivalent}$ yields the interval from -.20 to .97.

Had we tried to compute a 95% confidence interval around the obtained value of r_{sample} (i.e., 1.00, with a Z_r value of $+\ \infty$) we would have found it to show no uncertainty at all, a result that is entirely unreasonable, since the population correlation is not known to be 1.00 based on those six data points.

$r_{equivalent}$ vs. r_{sample}

In what sense is $r_{equivalent}$ a more accurate estimate of the population correlation than is the sample correlation, r_{sample}? A formal answer to this question is based on the fact that r_{sample}, although approximately unbiased for the population correlation, in small samples is a poor estimate. For example, suppose that in the population 80% of vaccinated animals survive while only 20% of unvaccinated animals survive. That difference in survival rates is associated with a correlation between vaccination and survival of .60. If we repeated our experiment on 3 vaccinated and 3 unvaccinated animals over and over, we would often find r_{sample} of 1.00 even though we know the population correlation is only .60. If the population survival rate for vaccinated animals were 90% while only 10% of unvaccinated animals survived, we would be even more likely to see r_{sample} values of 1.00, but our population value of r would still be far from 1.00; it would be .80. Even if 95% of vaccinated animals survived, while only 5% of unvaccinated animals survived, we would still have a population correlation of only .90 while obtaining r_{sample} values of 1.00 most of the time.

Insert Table 11 about here

Table 11 illustrates further that $r_{equivalent}$ based on exact p-values behaves in an intuitively more realistic way than r_{sample} in small samples. Table 11 shows the results of 8 hypothetical small to modest-sized studies of the effects of treatment on primate survival with *N*s ranging from 2 to 40. For each study, we report the p-value based on Fisher's exact test along with the associated $r_{equivalent}$ and the sample correlation, r_{sample}. As sample size, N, increases, the p-value decreases, and $r_{equivalent}$ increases; however, r_{sample} never changes -- it remains at 1.00.

Generality and Limitations of $r_{equivalent}$

The index $r_{equivalent}$ can be used in a wide variety of contexts beyond the simple contrasts computed among two or more treatment conditions. As long as a contrast is involved, comparisons among conditions leading to t tests or Z tests (or to F tests with 1 *df* in the numerator, or χ^2 tests on 1 *df*), can all be used to compute $r_{equivalent}$.

Although $r_{equivalent}$ is widely calculable, we emphasize that $r_{equivalent}$ is not a uniformly optimal procedure. It is not intended to be a kind of final common pathway effect size indicator. It is instead, designed specifically for those situations in which (a) the alternative is to have no effect size estimate at all (e.g., only sample sizes and *p*-values are known for a study), or (b) nonparametric procedures were employed for which there are no currently accepted effect size indicators, or (c) sample sizes are so small or data so nonnormal that the directly computed effect sizes would be more misleading than the computed value of $r_{equivalent}$.

To conclude with a medical analogy: we think of $r_{equivalent}$ as a first aid kit to be used for the time being until we can get to a highly sophisticated medical center. The medical center would be better, but it may be a long way away.

To come now to a close: It's been a long discussion of correlations, contrasts, and conceptual clarity. But I hope that some of what's been presented here persuades you to join me in appreciation of what Pearson has done for us and is still doing for us. Three cheers for Pearson's *r*!

References

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.). *Handbook of research synthesis* (pp. 399-409). New York: Russell Sage Foundation.

Campbell, D. T., & Fiske, D. (1959). Convergent and disciminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, J. (1965). Some statistical issues in psychological research . In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw Hill.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*, 997-1003.

Cohen, P. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H. M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology*, *41*, 1013-1018.

Cronbach, L, & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics, 13*, 131-147.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.). *Handbook of research synthesis* (pp. 245-260). New York: Russell Sage Foundation.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin, 100*, 257-269.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Kenny, D.A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. Shrout & S. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 111-124). Mahwah, N.J.: Erlbaum.

Kolata, G. B. (1981). Drug found to help heart attack survivors. *Science, 214,* 774-775.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Meng, X.L., Rosenthal, R., & Rubin, D.B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Pearson, K. (1896). Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, series A, *187*, 253-318.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, Nov. 5, 1243-1246.

Reichardt, C., & Coleman, S.C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivariate Behavioral Research*, *30*, 513-538.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*, 1-30.

Rosenthal, R. (1991a). *Meta-analytic procedures for social research*. (Rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1991b). Some indices of the reliability of peer review. *Behavioral and Brain Sciences*, *14*, 160-161.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.). *Handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R. (2000). Effect sizes in behavioral and biomedical research: Estimation and interpretation. In L. Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy*. Vol. 1, (pp. 121-139). Thousand Oaks, CA: Sage.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rosenthal, R., & DiMatteo, M. R. (2002). Meta-analysis, pp. 391-428. In J. Wixted (Ed.) *Stevens' Handbook of Experimental Psychology* (3rd ed.). Volume IV (Methodology in Experimental Psychology), Wiley.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research* (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The behavioral and brain sciences*, *3*, 377-386.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*, 395-396.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166-169.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, *5*, 329-334.

Rosenthal, R., & Rubin, D. B. (1998). Some new effect sizes for tables of counts. Unpublished manuscript.

Rosenthal, R., & Rubin, D. B. (2002). $r_{equivalent}$: A general effect size indicator. Submitted for publication.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, *1*, 331-340.

Rosnow, R. L., & Rosenthal, R. (In press). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*.

Rosnow, R. L., Rosenthal. R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*, 446-453.

Shrout, P., & Fiske, S. (1995). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Hillsdale, N.J.: Erlbaum.

Siegel, S. , & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2^{nd} ed.). New York: McGraw-Hill.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318*, 262-264.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. *Journal of the American Statistical Association, 54*, 30-34.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Westen, D., & Rosenthal, R. (2002). Quantifying construct validity: Two simple measures. Manuscript submitted for publication.

Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wothke, W. (1995). Covariance components analysis of the multitrait-multimethod matrix. In P. Shrout & S. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125-144). Mahwah, N.J.: Erlbaum.

Author Notes

This paper was prepared for the Award Address for the Samuel J. Messick Distinguished Scientific Contributions Award of APA Division 5: Evaluation, Measurement, and Statistics, given on August 22, 2002, at the meeting of the American Psychological Association held in Chicago.

Correspondence concerning this paper should be addressed to Robert Rosenthal, Department of Psychology, University of California, Riverside, Riverside, CA 92521-0426.

Table 1

Karl Pearson's (1904) Meta-Analysis: Correlations Between Smallpox Vaccination and Survival

Study		Pearson r
1		.60
2		.66
3		.77
4		.58
5		.58
6		.63
Mean		.64
Median		.61
S		.072
Standard Error $(S/\sqrt{6})$		.029
95% Confidence Interval	From	.56
	To	.72
One sample $t_{(5)}$		21.68
p		.000002
r		.99

Note: Calculations were carried out on untransformed *r*s since Fisher's Z_r transformation had not yet been invented (and because of the homogeneity of the obtained *r*s).

Table 2

Interpretive Data for Results of Table 1

A. Binomial Effect Size Display of Mean *r*

	Live	Die	Σ
Vaccinated	82	18	100
Untreated	18	82	100
Σ	100	100	200

B. Counternull Value of Mean *r*
.91

C. Coefficient of Robustness
M/S = 8.89[a]

D. File Drawer Tolerance for Future Null Results
1,045[b]

[a] 5.52 when based on Z_r rather than r.

[b] 401 when based on Z_r rather than r.

Table 3

Examples of Percent Agreement

A. Two Cases of 98 Percent Agreement

Smith's Results

	Judge A	
Judge B	Frown	No frown
Frown	98	1
No frown	1	0

Agreement = 98%, but

$r_{AB} = -.01;\ \chi^2(1) = 0.01$

Jones's Results

	Judge C	
Judge D	Frown	No frown
Frown	49	1
No frown	1	49

Agreement = 98%, but

$r_{CD} = +.96;\ \chi^2(1) = 92.16$

B. Two Cases of 50 Percent Agreement

North's Results

	Judge E	
Judge F	Frown	No frown
Frown	50	25
No frown	25	0

Agreement = 50%, but

$r_{EF} = -.33;\ \chi^2_{(1)} = 11.11$

West's Results

	Judge G	
Judge H	Frown	No frown
Frown	25	50
No frown	0	25

Agreement = 50%, but

$r_{GH} = +.33;\ \chi^2_{(1)} = 11.11$

Table 4

Results of Two Diagnosticians' Classification of 100 Persons into One of Four Categories

		Judge 1				
		A Schizophrenic	B Neurotic	C Normal	D Brain-damaged	Σ
Judge 2	A Schizophrenic	13	0	0	12	25
	B Neurotic	0	12	13	0	25
	C Normal	0	13	12	0	25
	D Brain-damaged	12	0	0	13	25
	Σ	25	25	25	25	100

$$kappa(df = 9) = \frac{O-E}{N-E} = \frac{50-25}{100-25} = .333$$

Table 5

Breakdown of the 9 df Omnibus Table of Counts of Table 4 into Six Specific (Focused) Reliabilities of df = 1 Each.

	A	B	
	Schizophrenic	Neurotic	Σ
A Schizophrenic	13	0	13
B Neurotic	0	12	12
Σ	13	12	25

kappa = 1.00

	A	C	
	Schizophrenic	Normal	Σ
A Schizophrenic	13	0	13
C Normal	0	12	12
Σ	13	12	25

kappa = 1.00

	A	D	
	Schizophrenic	Brain-damaged	Σ
A Schizophrenic	13	12	25
D Brain-damaged	12	13	25
Σ	25	25	50

kappa = .04

	B	C	
	Neurotic	Normal	Σ
B Neurotic	12	13	25
C Normal	13	12	25
Σ	25	25	50

kappa = -.04

	B	D	
	Neurotic	Brain-damaged	Σ
B Neurotic	12	0	12
D Brain-damaged	0	13	13
Σ	12	13	25

kappa = 1.00

	C	D	
	Normal	Brain-damaged	Σ
C Normal	12	0	12
D Brain-damaged	0	13	13
Σ	12	13	25

kappa = 1.00

Table 6

Alternative Results of Two Diagnosticians' Classification of 100 Persons into One of Four Categories

		Judge 1				
		A	B	C	D	Σ
	A	25	0	0	0	25
Judge 2	B	0	0	25	0	25
	C	0	25	0	0	25
	D	0	0	0	25	25
	Σ	25	25	25	25	100

$$kappa(df = 9) = \frac{O-E}{N-E} = \frac{50-25}{100-25} = .333$$

Table 7

Breakdown of the 9 df Omnibus Table of Counts of Table 6 into Six Specific (Focused) Reliabilities of df = 1 Each

	A	B	Σ
A	25	0	25
B	0	0	0
Σ	25	0	25

kappa = .00

	A	C	Σ
A	25	0	25
C	0	0	0
Σ	25	0	25

kappa = .00

	A	D	Σ
A	25	0	25
D	0	25	25
Σ	25	25	50

kappa = 1.00

	B	C	Σ
B	0	25	25
C	25	0	25
Σ	25	25	50

kappa = -1.00

	B	D	Σ
B	0	0	0
D	0	25	25
Σ	0	25	25

kappa = .00

	C	D	Σ
C	0	0	0
D	0	25	25
Σ	0	25	25

kappa = .00

Table 8

Three Examples of Four Effect Size Estimates

	Die	Live	Σ	1 Relative Risk $\left(\frac{A}{A+B} \Big/ \frac{C}{C+D}\right)$	2 Odds Ratio $\left(\frac{A}{B} \Big/ \frac{C}{D}\right)$	3 Risk Difference $\left(\frac{A}{A+B} - \frac{C}{C+D}\right)$	4 r^a
Control	A	B	A + B				
Treatment	C	D	C + D				
Σ	A + C	B + D	N				
Study I							
	Die	Live	Σ				
Control	10	990	1,000	10.00	10.09	.01	.06
Treatment	1	999	1,000				
Σ	11	1,989	2,000				
Study II							
	Die	Live	Σ				
Control	10	10	20	10.00	19.00	.45	.50
Treatment	1	19	20				
Σ	11	29	40				
Study III							
	Die	Live	Σ				
Control	10	0	10	10.00	∞	.90	.90
Treatment	1	9	10				
Σ	11	9	20				

[a] $$\frac{(AD-BC)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Table 9

Standardized Outcomes of Table 8

				1	2	3
	Die	Live	Σ	Standardized Relative Risk	Standardized Odds Ratio	Standardized Risk Difference (*r*)
Control	A	C	A + C	(A/C)	$(A/C)^2$	(A-C)/100
Treatment	C	A	A + C			
Σ	A + C	A + C	N			
<u>Study I</u>						
	Die	Live	Σ			
Control	53	47	100	1.13	1.27	.06
Treatment	47	53	100			
Σ	100	100	200			
<u>Study II</u>						
	Die	Live	Σ			
Control	75	25	100	3.00	9.00	.50
Treatment	25	75	100			
Σ	100	100	200			
<u>Study III</u>						
	Die	Live	Σ			
Control	95	5	100	19.00	361.00	.90
Treatment	5	95	100			
Σ	100	100	200			

Table 10

Correlations Between a New Measure of Interpersonal Skill and Four Other Measues (N=15)

Other Measures	New Measure (*Y*)	Verbal IQ (A)	Nonverbal Decoding (B)	Agreeableness (C)
A Verbal IQ	.63	-------		
B Nonverbal Decoding	.53	.38	-------	
C Agreeableness	.54	.36	.38	-------
D Conscientiousness	-.03	-.19	.12	.60

Note: Contrast weights for Measures A, B, C, D are +1, +1, +1, -3, respectively, based on predicted correlations with the new measure of +.50, +.50, +.50, +.10.

Table 11

Results of Eight Studies Showing *N*, *p*, $r_{equivalent}$, and r_{sample}

Study		Results		*N*	One-tailed exact *p*	$r_{equivalent}$	r_{sample}
		Survive	Die				
1	Treatment	1	0	2	.50	.00	1.00
	Control	0	1				
		Survive	Die				
2	Treatment	2	0	3	.33	.50	1.00
	Control	0	1				
		Survive	Die				
3	Treatment	2	0	4	.17	.67	1.00
	Control	0	2				
		Survive	Die				
4	Treatment	3	0	5	.10	.69	1.00
	Control	0	2				
		Survive	Die				
5	Treatment	3	0	6	.050	.73	1.00
	Control	0	3				
		Survive	Die				
6	Treatment	5	0	10	.0040	.78	1.00
	Control	0	5				
		Survive	Die				
7	Treatment	10	0	20	.0000054	.82	1.00
	Control	0	10				
		Survive	Die				
8	Treatment	20	0	40	$7.25/10^{12}$	.84	1.00
	Control	0	20				

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Correlations, Contrasts, and Conceptual Clarity

Author(s): Robert Rosenthal

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to each document.

If permission is granted to reproduce and disseminate the identified documents, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Robert Rosenthal TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate these documents as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Robert Rosenthal	Printed Name/Position/Title: ROBERT ROSENTHAL DISTINGUISHED PROFESSOR	
Organization/Address: DEPT. OF PSYCHOLOGY, UCR RIVERSIDE, CA 92521-0426	Telephone: 909 787 4503	FAX: 909 789 9269
	E-Mail Address:	Date: 12/12/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of these documents from another source, please provide the following information regarding the availability of these documents. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Counseling & Student Services
University of North Carolina at Greensboro
201 Ferguson Building
PO Box 26171
Greensboro, NC 27402-6171